Self-averaging and On-line Learning

G. Reents and R. Urbanczik
Institut für Theoretische Physik
Universität Würzburg
Am Hubland
D-97074 Würzburg
Germany

January 1998

(to appear in Phys. Rev. Letters)

Abstract

Conditions are given under which one may prove that the stochastic dynamics of on-line learning can be described by the deterministic evolution of a finite set of order parameters in the thermodynamic limit. A global constraint on the average magnitude of the increments in the stochastic process is necessary to ensure self-averaging. In the absence of such a constraint, convergence may only be in probability.

On-line learning, introduced in [1, 2], has become an important paradigm in the analysis of neural networks. Not only has it enabled the understanding of specific algorithms for a wide range of supervised learning scenarios and network architectures, e.g. [3, 4, 5], but one may also derive learning algorithms which are highly optimized for a specific problem, e.g. [6, 7]. Furthermore it is also possible to analyze unsupervised learning within this framework [8, 9].

The key assumption in on-line learning is that the adaption of the network is driven, at each time step, by the presentation of a single pattern which is picked independently of any previous patterns. Thus the evolution of the state vector of the network is governed by a stochastic (Markov) process. However, if the underlying distribution of patterns is not too complicated, it is possible to characterize the performance of the network by a few order

parameters, and one expects these parameters to be self-averaging for large networks. This makes it possible to map the stochastic evolution of the state vector onto a deterministic evolution of the order parameters, thus greatly facilitating a theoretical understanding.

While the self-averaging properties of the order parameters may usually be well observed numerically, a rigorous proof of this crucial fact has so far been lacking. The goal of this paper is to give conditions on the stochastic dynamics which ensure that it may be described by deterministic order parameters in the thermodynamic limit and to clarify the nature of the convergence. We first review the customary heuristic derivation of the deterministic equations in the context of the perceptron learning rule. Next, a framework for the analysis of on-line learning is established which is general enough to cover many of the scenarios discussed in the literature. Within this framework we prove convergence by exploiting the fact that the thermodynamic limit is in some ways analogous to a small step size limit in order parameter space. (The small step size limit in weight space has been considered in [10]). Of course some assumptions about the stochastic process are needed for the proof, and the concluding paragraphs discuss examples to show that these assumptions, while not being necessary for convergence, are nevertheless quite reasonable. [11]

To fix ideas let us first consider the perceptron learning rule. We are given a sequence of inputs $\xi^\mu \in \mathbb{R}^N$ and corresponding outputs $s^\mu \in \{-1, 1\}$ and we assume that the inputs are picked independently from some probability distribution. We hope to approximate the mapping from input to output by a perceptron s_w which implements the function $s_w(\xi) = \mathrm{sign}(w^T\xi)$ for $w, \xi \in \mathbb{R}^N$. To this end, we use a new example (ξ^μ, s^μ) to update our current estimate of a good weight vector w^μ by

$$w^{\mu+1} = w^\mu + \frac{\eta}{N}\frac{s^\mu - s_w(\xi^\mu)}{2}\xi^\mu \ , \tag{1}$$

where $\eta \equiv \eta(\mu/N)$ is a possibly time dependent learning rate. For simplicity we assume that the output is itself given by a perceptron with weight vector B, $s^\mu = s_B(\xi^\mu)$. Then the quantity of interest is the angle between w^μ and B which may be calculated from the overlaps $r^\mu = B^T w^\mu$ and $q^\mu = w^{\mu T} w^\mu$. One easily finds recursive equations for $r^{\mu+1}$ and $q^{\mu+1}$ using (1). These will of course still depend on the entire input sequence $\{\xi^\mu\}$ but if we assume that the input components are picked independently from the normal distribution,

it is straightforward to average over the last input and find:

$$\begin{aligned}\langle r^{\mu+1}\rangle_{\xi^\mu} &= r^\mu + \frac{\eta}{N}\frac{1-r^\mu/\sqrt{q^\mu}}{\sqrt{2\pi}} \\ \langle q^{\mu+1}\rangle_{\xi^\mu} &= q^\mu + \frac{2\eta}{N}\frac{r^\mu-\sqrt{q^\mu}}{\sqrt{2\pi}} + \frac{\eta^2}{N}\frac{\arccos\left(r^\mu/\sqrt{q^\mu}\right)}{\pi} \ . \end{aligned} \tag{2}$$

Since r^μ and q^μ are still stochastic quantities the above equations do not seem very helpful. What one would really like to calculate, is averages over the entire sequence of inputs up to time μ, for instance:

$$\langle r^\mu\rangle_\mu = \langle r^\mu\rangle_{\xi^0,\xi^1,\ldots,\xi^{\mu-1}} \ . \tag{3}$$

At this point it is customary to argue that in the thermodynamic limit, $N\to\infty$ but $\mu=\mathcal{O}(N)$, the overlaps will be self-averaging and that r^μ will thus be close to $\langle r^\mu\rangle_\mu$ for large N. Substituting the averages $(\langle r^\mu\rangle_\mu, \langle q^\mu\rangle_\mu)$ for the stochastic quantities (r^μ, q^μ) in the iteration (2), leads to deterministic finite difference equations, and, taking the large N limit once again, one arrives at the set of differential equations:

$$\begin{aligned}\dot r &= \eta(t)\frac{1-r/\sqrt{q}}{\sqrt{2\pi}} \\ \dot q &= 2\eta(t)\frac{r-\sqrt{q}}{\sqrt{2\pi}} + \eta(t)^2\frac{\arccos\left(r/\sqrt{q}\right)}{\pi} \ . \end{aligned} \tag{4}$$

One now claims that for large N and identical or similar initial conditions (r^{tN}, q^{tN}) will be close to $(r(t), q(t))$.

The main goal of this paper is to make such claims precise and give conditions under which they are rigorously true. Since we do not want to confine ourselves to the perceptron, the theory should e.g. cover learning in multilayer perceptrons as well, we introduce a somewhat more general framework. Consider an iteration of the form

$$J^{\mu+1} = J^\mu + N^{-1} f(J^\mu, \xi^\mu) \ , \tag{5}$$

where the patterns ξ^μ are picked independently from a probability distribution on $\mathbb{R}^{L_N}$. The state vectors J^μ and the increments lie in some $\mathbb{R}^{M_N}$. The standard case in on-line learning is that the input dimension L_N and the system size M_N are on the order of N. Further let

$$\mathcal{Q} : \mathbb{R}^{M_N}\times\mathbb{R}^{M_N} \to V \subset \mathbb{R}^n \tag{6}$$

be a symmetric, bilinear mapping. The intended interpretation is that $\mathcal{Q}(J^\mu, J^\mu)$ gives the order parameters of the problem.

It might seem that some order parameters, such as r^μ in the above perceptron rule, cannot be obtained by applying a quadratic form to the state vector. However, by using a larger state vector this can always be achieved, as well as for instance the transformation of a nonautonomous system into an autonomous one. In the perceptron case one may formally augment equation (1) with the following set of equations:

$$b^{\mu+1} = b^\mu, \quad \tau_1^{\mu+1} = \tau_1^\mu + 1/N, \quad \tau_2^{\mu+1} = \tau_2^\mu \tag{7}$$

and fix the initial conditions for the new recursions by $b^0 = B,\ \tau_1^0 = 0,\ \tau_2^0 = 1$. By aggregating w, b, τ_1, τ_2 to form a vector J of dimension $2N + 2$ the set of equations (1,7) is of the general form (5). Furthermore we define a bilinear symmetric form taking values in $\mathbb{R}^3$ via $\mathcal{Q}(J, \hat{J}) = (w^T\hat{w}, b^T\hat{w}, \tau_1\hat{\tau_2})$, where $\hat{J} = (\hat{w}, \hat{b}, \hat{\tau}_1, \hat{\tau}_2)$. Now $\mathcal{Q}(J^\mu, J^\mu) = (q^\mu, r^\mu, \mu/N)$ and since we thus obtain the order parameters of the perceptron rule (2), this rule is indeed just a special case of the general framework (5,6).

Before proceeding with the general theory, a word of caution regarding our notation is in order. We are of course not considering a single stochastic process but a sequence of these. But to reduce notational overhead we have suppressed the index N in symbols such as $J^\mu, \xi^\mu, f, \mathcal{Q}$ and the factor N^{-1} in (5) is just an attempt at suggestive notation. It is, however, crucial that the number n of order parameters (6) and the set of their possible values V be independent of N.

Writing $\mathcal{Q}(J, \hat{J})$ more conveniently as $J * \hat{J}$, the iteration (5) yields the following relation for the order parameters $Q^\mu = J^\mu * J^\mu$:

$$\begin{aligned} Q^{\mu+1} &= Q^\mu + N^{-1}F(J^\mu, \xi^\mu) \ , & (8)\\ F(J, \xi) &= 2J * f(J, \xi) + N^{-1}f(J, \xi) * f(J, \xi) \ . & (9) \end{aligned}$$

For the Q^μ to be the order parameters of the problem, the input average of the increment function $F(J, \xi)$ should for large N converge to a quantity which only depends on $J * J$. At this point we shall just write

$$\langle F(J, \xi)\rangle_\xi \to G(J * J) \tag{10}$$

and be more precise about the kind of convergence later. That the limit $N \to \infty$ is not just the limit of small step size but a thermodynamic one, in which the system size M_N and the input size L_N may diverge, is important

in the definition of G: The N^{-1} term in the increment function F given by (9) will in general give a finite contribution to G.

We may now associate to the stochastic process (8) the deterministic trajectory

$$\dot{Q} = G(Q) \tag{11}$$

and ask whether Q^{tN} converges to $Q(t)$ for large N. Here and in the sequel, the convention is that a real expression is rounded when it appears in the position of an integer index, like tN in Q^{tN}. For a given initial condition $Q(0)$, we assume that (11) has a solution Q(t) up to certain time α. We further require the existence of a compact set U which contains a neighborhood of the trajectory, more formally

$$\{x \in V : |x - Q(t)| \leq \epsilon\} \subset U \ . \tag{12}$$

Note that here and in the sequel we assume $0 \leq t \leq \alpha$.

We are now in the position to state conditions for the convergence of the stochastic process to the deterministic trajectory:

(a) $|G(Q_1) - G(Q_0)| < C|Q_1 - Q_0|$ for $Q_1, Q_0 \in U$ and some constant C.

(b) $|Q^0 - Q(0)| < l(N)$,
$|\langle F(J,\xi)\rangle_\xi - G(J * J)| < h(N)$ if $J * J \in U$,
for suitable functions l and h with $\lim_{N\to\infty} l(N) = \lim_{N\to\infty} h(N) = 0$.

(c) $\langle |F(J,\xi)|^2\rangle_\xi < C^2(|J*J|+1)^2$, for convenience we use the same constant C, independent of N as in condition (a).

The Lipschitz condition (a) makes sure that there is a unique deterministic trajectory given the initial value $Q(0)$. Note that this condition is only required to hold in the neighborhood U of the deterministic trajectory. Indeed, by considering e.g. the limit $q \to 0$ for the perceptron rule (4), one sees that even for this simple case no global Lipschitz condition holds. Condition (b) clarifies the required relationship between the stochastic process and the deterministic trajectory: Initial conditions should converge and so should the increments, at least on average and in the neighborhood U. The perhaps most interesting condition is (c). In the case of the perceptron learning rule, the fourth moments of the input distribution must exist, for the LHS of (c) to be defined. Further the condition implies

(c') $|\langle F(J,\xi)\rangle_\xi| < C(|J*J|+1)$.

and is thus a global constraint on the growth of the increments.

Given these conditions, one may prove convergence by considering the difference $\Delta^\mu = Q^\mu - Q(\mu/N)$ between the stochastic and the deterministic trajectory. Using the abbreviation $g^\mu = N(Q(\mu/N + 1/N) - Q(\mu/N))$ the following recursive relation for the variance σ^μ of Δ^μ is obtained from (8):

$$\begin{aligned}
\sigma^{\mu+1} &= \left\langle |\Delta^{\mu+1}|^2 \right\rangle_{\mu+1} \\
&= \left\langle |\Delta^\mu + \frac{1}{N}(F(J^\mu, \xi^\mu) - g^\mu)|^2 \right\rangle_{\mu+1} \\
&= \sigma^\mu + \\
&\quad 2/N \left\langle \Delta^{\mu T}(F(J^\mu, \xi^\mu) - g^\mu) \right\rangle_{\mu+1} + \\
&\quad 1/N^2 \left\langle |F(J^\mu, \xi^\mu)|^2 - 2F(J^\mu, \xi^\mu)^T g^\mu + |g^\mu|^2 \right\rangle_{\mu+1}
\end{aligned} \tag{13}$$

The next step is to find an upper bound on the increments to σ^μ which depends only on σ^μ itself. For the $2/N$-term in the above equation we need to distinguish between Q^μ being in U or not. So we rewrite this term as:

$$\begin{aligned}
&\left\langle \Delta^{\mu T}(F(J^\mu, \xi^\mu) - g^\mu) \right\rangle_{\mu+1} = \\
&\quad \left\langle (1 - \theta(|\Delta^\mu| - \epsilon))\Delta^{\mu T}(\langle F(J^\mu, \xi)\rangle_\xi - g^\mu) \right\rangle_\mu + \\
&\quad \left\langle \theta(|\Delta^\mu| - \epsilon)\Delta^{\mu T}(\langle F(J^\mu, \xi)\rangle_\xi - g^\mu) \right\rangle_\mu
\end{aligned} \tag{14}$$

In the first summand one rewrites the difference as:

$$\begin{aligned}
\langle F(J^\mu, \xi)\rangle_\xi - g^\mu &= \langle F(J^\mu, \xi)\rangle_\xi - G(Q^\mu) + \\
&\quad G(Q^\mu) - G(Q(\mu/N)) + \\
&\quad G(Q(\mu/N)) - g^\mu \, .
\end{aligned} \tag{15}$$

Expanding the product of $\Delta^{\mu T}$ with the above RHS, applying the triangle inequality and then (b) and (a) one obtains [12] an upper bound:

$$\begin{aligned}
&\left\langle (1 - \theta(|\Delta^\mu| - \epsilon))\Delta^{\mu T}(\langle F(J^\mu, \xi)\rangle_\xi - g^\mu) \right\rangle_\mu \leq \\
&\qquad (h(N) + C^2/N)\sqrt{\sigma^\mu} + C\sigma^\mu.
\end{aligned} \tag{16}$$

To bound the second term in (14) one uses that the growth condition (c) implies

$$|\Delta^{\mu T}(\langle F(J^\mu, \xi)\rangle_\xi - g^\mu)| \leq |\Delta^\mu|^2 C + |\Delta^\mu|(C^2 + C) \tag{17}$$

and thus:

$$\left\langle \theta(|\Delta^\mu| - \epsilon)\Delta^{\mu T}(\langle F(J^\mu, \xi)\rangle_\xi - g^\mu)\right\rangle_\mu \leq$$
$$C\sigma^\mu + (C^2 + C)\left\langle \theta(|\Delta^\mu| - \epsilon)|\Delta^\mu|\right\rangle_\mu \ . \tag{18}$$

The remaining average can be further simplified by applying a Tschebyscheff inequality: $\langle \theta(|\Delta^\mu| - \epsilon)|\Delta^\mu|\rangle_\mu \leq \sigma^\mu/\epsilon$. Using the growth condition (c) one may bound the $1/N^2$ term in (13) and combining this with (16,18) finally yields:

$$\begin{aligned} A/N &\geq (\sigma^{\mu+1} - \sigma^\mu)/u\,(\sigma^\mu)\ , \\ u(\sigma) &= (h(N)\sqrt{\sigma} + \sigma) + N^{-1}(1 + \sqrt{\sigma} + \sigma)\ , \end{aligned} \tag{19}$$

for a suitable positive A which depends only on C and ϵ. Note that the bound (19) holds for any μ and N. By rewriting its RHS as an integral and summing over μ one finds $\int_{\sigma^0}^{\sigma^\mu} u(\sigma)^{-1}\mathrm{d}\sigma \leq A\mu/N$. Replacing the term $\sqrt{\sigma}$ in $u(\sigma)$ by its value at the upper limit σ^μ, makes the integral both smaller and simpler and in the end yields the following key inequality:

$$\sigma^\mu \leq 4(N^{-1} + l(N)^2 + h(N)^2)\exp\left(4A\frac{\mu}{N}\right)\ . \tag{20}$$

Consequently for $\mu = tN$ the variance decays to zero in the large N limit, the probability of Q^{tN} deviating from the sequence average $\langle Q^{tN}\rangle_{tN}$ vanishes and the stochastic process is self-averaging. [13]

Let us next consider relaxing the global constraint (c). Assume a situation where (c) holds for $J * J \in U$ but not necessarily outside of U. We may then replace the update rule f in (5) by

$$\tilde{f}(J, \xi) = \begin{cases} f(J, \xi) & \text{if } \ J * J \in U \\ 0 & \text{else.} \end{cases} \tag{21}$$

Then, for identical initial conditions, the deterministic trajectory of f and $\tilde{f}$ will be the same. Further all of the conditions (a-c) hold for $\tilde{f}$ and this stochastic process is self-averaging. Since the deterministic trajectory lies strictly in the interior of U and since the increments $\tilde{f}$ are zero outside of U, this implies that the probability of $\tilde{Q}^{tN}$ not lying in U (for any $t \in [0, \alpha]$) must vanish for large N. So given the same input sequence $\tilde{f}$ will typically give the same result as the unmodified dynamics f, and thus Q^{tN} converges to $Q(t)$ in probability. Thus, for this weaker notion of convergence, no global constraint is needed.

To be able to conclude, however, that in such a situation the stochastic process given by f is self-averaging, we would have to know that Q^{tN} is well behaved on untypical sequences as well. A simple example will be sufficient to show that this need not be the case. Consider the following random walk with a step size which depends on the length of the current vector:

$$J^{\mu+1} = J^\mu + N^{-1}(Q^\mu - 1)\xi^\mu. \tag{22}$$

Here we assume $J^\mu, \xi^\mu \in \mathbb{R}^N$, $Q^\mu = |J^\mu|^2$ and the components of the ξ^μ are picked independently from the normal distribution. Averaging the self-overlap $Q^{\mu+1}$ with respect to the last input yields

$$\langle Q^{\mu+1}\rangle_{\xi^\mu} = Q^\mu + N^{-1}(Q^\mu - 1)^2 \tag{23}$$

and condition (c) is violated. The deterministic trajectory is given by $\dot{Q} = (Q-1)^2$ and while it is defined for all times if $Q(0) \leq 1$, it will diverge at some finite time if the initial condition has $Q(0) > 1$. Consequently one will expect Q^{tN} to diverge with N if Q^0 is greater than 1 and t is sufficiently large. To obtain a lower bound on this divergence, first note that by convexity of the RHS in (23)

$$\langle Q^{\mu+1}\rangle_{\mu+1} \geq \langle Q^\mu\rangle_\mu + N^{-1}(\langle Q^\mu\rangle_\mu - 1)^2 \,. \tag{24}$$

Setting $\tilde{Q}^\mu = \langle Q^\mu\rangle_\mu - 1$ yields $\tilde{Q}^{\mu+1} \geq \tilde{Q}^\mu + N^{-1}(\tilde{Q}^\mu)^2$ and dropping the first summand allows us to solve the recurrence and find

$$\tilde{Q}^\mu \geq (\tilde{Q}^{\mu_0}/\sqrt{N})^{2^{\mu-\mu_0}} \,. \tag{25}$$

Thus $\tilde{Q}^\mu$ (and $\langle Q^\mu\rangle_\mu$) will increase super-exponentially with μ if ever $\tilde{Q}^{\mu_0}$ becomes larger than $\sqrt{N}$.

Let us now consider the dynamics for an initial condition $Q^0 = Q(0) = 0$. By the general theory presented above Q^{tN} will with increasing N converge in probability to $Q(tN)$ for any fixed t. There is, however, a small probability of making a large first step. In particular, the probability of having $Q^1 > N$ is larger than $\exp(-P(N))$, where $P(N)$ is a suitable polynomial in N. Whenever such a rare event ($Q^1 > N$) occurs, due to (25) the following steps lead to an extremely fast growth. Consequently $\langle Q^{tN}\rangle_{tN}$ diverges with N for any positive t and the stochastic dynamics is not self-averaging in the thermodynamic limit.

While we believe that the conditions imposed on the stochastic process are not overly restrictive, they are not necessary for self-averaging to hold.

A case in point is the perceptron rule (1). While conditions (a-c) are true for an initial condition with $q(0) > 0$, the Lipschitz condition (a) is violated for $q(0) = 0$ and it is not possible to define $r/\sqrt{q}$ in a manner that would make the RHS of (4) continuous in the point $q(0) = 0$. Nevertheless, numerical simulations indicate that the perceptron rule is self-averaging for this initial condition. But this property is highly dependent on fine details of the input distribution: Instead of always choosing Gaussian inputs, consider presenting a Gaussian input only in one half of the steps and else presenting some fixed vector. More formally, let the input ξ be the random variable $\xi = \chi\tilde{\xi} + (1-\chi)N^{-1}b$, where $\tilde{\xi}$ has normally distributed components, χ is 0 or 1 with equal probability and b is a fixed vector with $|b| \leq 1$. The deterministic trajectory does not depend on the specific choice of b and is, up to a factor of $1/2$, still given by (4). If we choose $b = 0$, the stochastic process is essentially the same as for plain Gaussian inputs, except that, on average, the weight vector does not change in every second step. But now consider the choice $b = B$ and assume that $\mathrm{sign}(0) = 0$. For $q^0 = 0$ in one half of the cases perfect generalization will be achieved in the first step and subsequently the weight vector will not change. However, if the initial input is Gaussian ($\chi^0 = 1$), any subsequent presentation of B as input will not change the weight vector since we already have positive overlap with the teacher and the subsequent dynamics will be the same as for the choice $b = 0$. So, for $b = B$, the first step is crucial and the on-line dynamics is not self-averaging.

While the above example is rather construed, it does nevertheless show that the self-averaging properties of on-line learning can depend on rather minute details of the stochastic process if the Lipschitz condition is violated. Consequently we believe that it is difficult to find easily verifiable conditions for self-averaging which are much weaker than the ones presented here.

The authors wish to acknowledge helpful discussions with M. Biehl, W. Kinzel, and M. Opper. The work of one of us (R.U.) was supported by the *Deutsche Forschungsgemeinschaft*.

References

[1] S. Amari, IEEE **EC-16**, 299, 1967.

[2] W. Kinzel and P. Ruján, Europhys. Lett., **13**, 473, 1990.

[3] M. Biehl and H. Schwarze, J. Phys. A, **26**, 2651, 1993.

[4] N. Barkai, H.S. Seung, and H. Sompolinsky, Phys. Rev. Lett., **75**, 1415, 1995.

[5] D. Saad and S. Solla, Phys. Rev. Lett., **74**, 4337, 1995.

[6] O. Kinouchi and N. Caticha, J. Phys. A, **25**, 6243, 1992.

[7] M. Opper, Phys. Rev. Lett., **77**, 4671, 1996.

[8] M. Biehl, Europhys. Lett., **25**, 391, 1994.

[9] C. Van den Broeck and P. Reimann, Phys. Rev. Lett., **76**, 2188, 1996.

[10] T.M. Heskes and B. Kappen, in *Mathematical approaches to neural networks*, edited by J.G. Taylor, (North-Holland, Amsterdam, 1993).

[11] Indeed, our convergence conditions apply to most of the on-line scenarios discussed in the literature. To be specific, they hold e. g. for backpropagation in MLP's as discussed in [5], assuming Gaussian inputs and exluding the pathological case of a vanishing initial weight vector.

[12] Here and in the sequel we also assume that C is large enough for the following conditions to hold: $|g^\mu - G(Q(\mu/N))| \leq C^2/N$, $|g^\mu| \leq C$ and $|Q(\mu/N)| + 1 < C$. That such a C can always be found, is a consequence of standard results in the theory of differential equations and the fact that G is Lipschitz on the compact set U.

[13] Further, any quantity which is a function of the Q^{tN} (e.g. an angle between vectors) will be self-averaging as well, if the function is Lipschitz. In case the function is only continuous, convergence in probability is still assured.